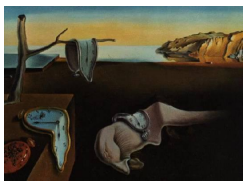

Journal of Statistical Software
March 2012, Volume 46, Issue 13. http://www.jstatsoft.org/

survivalBIV: Estimation of the Bivariate Distribution Function for Sequentially Ordered Events Under Univariate Censoring

Ana Moreira
University of Minho

Luís Meira-Machado
University of Minho

Abstract

In many medical studies, patients can experience several events. The times between consecutive events (gap times) are often of interest and lead to problems that have received much attention recently. In this work we consider the estimation of the bivariate distribution function for censored gap times, using **survivalBIV** a software application for R. Some related problems such as the estimation of the marginal distribution of the second gap time is also discussed. It describes the capabilities of the program for estimating these quantities using four different approaches, all using the Kaplan-Meier estimator of survival. One of these estimators is based on Bayes' theorem and Kaplan-Meier survival function. Two estimators were recently proposed using the Kaplan-Meier estimator pertaining to the distribution of the total time to weight the bivariate data (de Uña-Álvarez and Meira-Machado 2008 and de Uña-Álvarez and Amorim 2011). The software can also be used to implement the estimator proposed in Lin, Sun, and Ying (1999), which is based on inverse probability of censoring weighted. The software is illustrated using data from a bladder cancer study.

Keywords: censoring, Kaplan-Meier, multi-state model, gap times, inverse censoring.

1. Introduction

In longitudinal studies of disease, patients may experience several events through a follow-up period. In these studies, the sequentially ordered events (and the gap times) are often of interest. The events of concern can be of the same nature (e.g., cancer patients can experience recurrent disease episodes) or represent different states in the disease process (e.g., alive and disease-free, alive with recurrence and dead). If the events are of the same nature, this is usually referred as recurrent events, whereas if they represent different states they are

usually modeled through their intensity functions (Andersen, Borgan, Gill, and Keiding 1993; Hougaard 2000; Meira-Machado, de Uña-Álvarez, Cadarso-Suárez, and Andersen 2009).

Let (T_1, T_2) be a pair of gap times of successive events, which are observed subjected to random right-censoring. Let C be the right-censoring variable, assumed to be independent of (T_1, T_2) and let $Y = T_1 + T_2$ be the total time. Because of this, we only observe $(\widetilde{T}_{1i}, \widetilde{T}_{2i}, \Delta_{1i}, \Delta_{2i})$, $1 \leq i \leq n$, which are n independent replications of $(\widetilde{T}_1, \widetilde{T}_2, \Delta_1, \Delta_2)$, where $\widetilde{T}_1 = T_1 \wedge C$, $\Delta_1 = I(T_1 \leq C)$, and $\widetilde{T}_2 = T_2 \wedge C_2$, $\Delta_2 = I(T_2 \leq C_2)$ with $C_2 = (C - T_1)I(T_1 \leq C)$ the censoring variable of the second gap time. Define $\widetilde{Y} = Y \wedge C$ and let F_1 denote the distribution function of T_1, and G the survival function of the censoring time variable, C. Since T_1 and C are independent, the Kaplan-Meier product-limit estimator based on the pairs $(\widetilde{T}_{1i}, \Delta_{1i})$'s , consistently estimates the distribution F_1. Similarly, the distribution of the total time may be consistently estimated by the Kaplan-Meier estimator based on $(\widetilde{T}_{1i} + \widetilde{T}_{2i}, \Delta_{2i})$'s. Because T_2 and C_2 will be generally dependent, the estimation of the marginal distribution for the second gap time is not a simple issue. The same applies to the bivariate distribution function $F_{12}(x, y) = P(T_1 \leq x, T_2 \leq y)$. This issue has received much attention recently. Among others, it was investigated by Lin *et al.* (1999), Laan, Hubbard, and Robins (2002), Keilegom (2004), de Uña-Álvarez and Meira-Machado (2008) or de Uña-Álvarez and Amorim (2011).

In this work we present four methods (estimators) for the bivariate distribution function of the gap times. One simple estimator is based on Bayes' theorem and Kaplan-Meier survival function. This estimator is related to that proposed in Lin *et al.* (1999) and with estimators proposed by de Uña-Álvarez (de Uña-Álvarez and Meira-Machado 2008; de Uña-Álvarez and Amorim 2011) since all use (in different ways) the Kaplan-Meier estimator (Kaplan and Meier 1958). The estimator proposed by Lin in 1999 uses inverse probability of censoring weighted (`IPCW`) based on the Kaplan-Meier estimator. On the other hand, the idea behind both estimators proposed by de Uña-Álvarez is the use of the Kaplan-Meier estimator pertaining to the distribution of the total time to weight the bivariate data. The difference between these two methods is that the more recent paper uses a presmoothed version of the Kaplan-Meier estimator (Dikta 1998). Without smoothing, the estimator described in de Uña-Álvarez and Amorim (2011) reduces to that in the de Uña-Álvarez and Meira-Machado (2008).

This paper describes the R-based package **survivalBIV** (available from the Comprehensive R Archive Network at `http://CRAN.R-project.org/package=survivalBIV`) and its capabilities for implementing nonparametric and semiparametric estimators for the bivariate distribution function for censored gap times. In this article we explain and illustrate how numerical and graphical output for all methods can be obtained using the **survivalBIV** package.

The following section provides a brief introduction to the methodological background. All four estimators for the bivariate distribution function and marginal distribution of the second gap time are presented. An overview of the use of **survivalBIV** is given in Section 3. In Section 4 we explain how the package can be used to simulate bivariate censored data and how to use the several functions in the package. An example of its application is given using data from a bladder cancer study in Section 5. A discussion is in Section 6.

2. Methodological background

In this section we will present four different methods for estimating the bivariate distribution function $F_{12}(x, y) = P(T_1 \leq x, T_2 \leq y)$, all using the Kaplan-Meier estimator of survival. Some

related problems such as estimation of the marginal distribution of the second gap time will also be discussed.

2.1. Conditional Kaplan-Meier estimator

A simple estimator for the bivariate distribution function of the gap times is based on Bayes' theorem and Kaplan-Meier survival function (conditional Kaplan-Meier, `CKM`). Since $F_{12}(x,y) = P(T_1 \leq x, T_2 \leq y) = P(T_2 \leq y | T_1 \leq x)P(T_1 \leq x)$ one simple estimator for the bivariate distribution is given by

$$\widehat{F}_{12}(x,y) = \widehat{F}_1(x)\widehat{F}_{KM}(y|T_1 \leq x, \Delta_1 = 1) \tag{1}$$

where $\widehat{F}_1(x)$ is the Kaplan-Meier product-limit estimator based on the pairs $(\widetilde{T}_{1i}, \Delta_{1i})$'s and $\widehat{F}_{KM}(y)$ is the Kaplan-Meier estimator based on the pairs $(\widetilde{T}_{2i}, \Delta_{2i})$'s. The $\widehat{F}_{KM}(y|T_1 \leq x, \Delta_1 = 1)$ is the conditional distribution function for the subset of $T_1 \leq x$ and $\Delta_1 = 1$ (the Kaplan-Meier estimator based on the pairs $(\widetilde{T}_{2i}, \Delta_{2i})$'s such that $\widetilde{T}_{1i} \leq x$ and $\Delta_{1i} = 1$).

2.2. Kaplan-Meier weighted estimator

Another simple estimator was recently proposed by de Uña-Álvarez and Meira-Machado (2008). The idea behind estimation is to use the Kaplan-Meier estimator pertaining to the distribution of the total time to weight the bivariate data. The proposed estimator (Kaplan-Meier weighted estimator, `KMW`) is given by

$$\widetilde{F}_{12}(x,y) = \sum_{i=1}^{n} W_i I(\widetilde{T}_{1i} \leq x, \widetilde{T}_{2i} \leq y) \tag{2}$$

where $W_i = \frac{\Delta_{2i}}{n-R_i+1} \prod_{j=1}^{i-1} \left[1 - \frac{\Delta_{2j}}{n-R_j+1}\right]$ is the Kaplan-Meier weight attached to $\widetilde{Y}_i$ when estimating the marginal distribution of Y from $(\widetilde{Y}_i, \Delta_{2i})$'s, and for which the ranks of the censored $\widetilde{Y}_i$'s, R_i, are higher than those for uncensored values in the case of ties.

2.3. Kaplan-Meier presmooth weighted estimator

Recently, de Uña-Álvarez and Amorim (2011) propose a modification of estimator (2) based on presmoothing (Dikta 1998), which allows for a variance reduction in the presence of censoring. Basically, this method uses a presmoothed version of the Kaplan-Meier estimator (see e.g., Dikta 1998 and references therein) pertaining to the distribution of the total time to weight the bivariate data. This is obtained by replacing the censoring indicator variables in the expression of the Kaplan-Meier weights by a smooth fit of a binary regression. This estimator (Kaplan-Meier presmooth weighted estimator, `KMPW`) is expressed as

$$\breve{F}_{12}(x,y) = \sum_{i=1}^{n} W_i^{\star} I(\widetilde{T}_{1i} \leq x, \widetilde{T}_{2i} \leq y) \tag{3}$$

where $W_i^{\star} = \frac{m(\widetilde{T}_{1i}, \widetilde{Y}_i)}{n-R_i+1} \prod_{j=1}^{i-1} \left[1 - \frac{m(\widetilde{T}_{1j}, \widetilde{Y}_j)}{n-R_j+1}\right]$ are the presmoothed Kaplan-Meier weights. Here, $m(x,y) = P(\Delta_2 = 1 | \widetilde{T}_1 = x, \widetilde{Y} = y, \Delta_1 = 1)$, belongs to a parametric (smooth) family of binary regression curves, e.g., logistic. Our package provides the results assuming that m denotes a logistic regression model (`KMPW`). In practice, we assume that $m(x,y) = m(x,y;\beta)$ where

β is a vector of parameters which typically will be computed by maximizing the conditional likelihood of the Δ_2's given $(\widetilde{T}_1, \widetilde{T}_2)$ for those with $\Delta_1 = 1$.

Note that, unlike (2), the `KMPW` can attach positive mass to pair of gap times with censored second gap time. However, both estimators (2) and (3) attach a zero weight to pairs of gap times with censored first gap time. In the limit case of no presmoothing, the estimator (3) reduces to (2). Conditions under which both estimators are consistent is fully discussed in papers by de Uña-Álvarez and Meira-Machado (2008) and de Uña-Álvarez and Amorim (2011). In the latter paper the authors compare the performance of the presmoothed (semiparametric) estimator with the purely nonparametric estimator (without presmoothing) and concluded that the presmoothed estimator improves efficiency in the multivariate setup of gap times.

2.4. Inverse probability of censoring weighted estimator

Another estimator for the bivariate distribution function was proposed by Lin *et al.* (1999). This estimator is based on inverse probability of censoring weighted (`IPCW`). The rationale behind `IPCW` is that each subject that is observed at time u is representative of $1/G(u)$ individuals that might have been observed if there was no censoring. Lin's estimator is expressed as

$$\overline{F}_{12}(x,y) = \overline{H}(x,0) - \overline{H}(x,y) \tag{4}$$

where

$$\overline{H}(x,y) = \frac{1}{n}\sum_{i=1}^{n} \frac{I(\widetilde{T}_{1i} \le x, \widetilde{T}_{2i} > y)}{\widehat{G}(\widetilde{T}_{1i} + y)}$$

The censoring survivor function G is typically unknown and needs to be replaced by an estimate. This can be obtained by reversing the role of T and C, using a Kaplan-Meier estimate $\hat{G}$ of the censoring survivor function, i.e., using an estimate based on the $(\widetilde{T}_{1i}, 1 - \Delta_{1i})$'s (for the first term in the right-hand side of Equation 4) or $(\widetilde{Y}_i, 1-\Delta_{2i})$'s (for the second term in the right-hand side of Equation 4). This is the simplest choice and was assumed by Lin *et al.* (1999). Other procedures for estimation of G are appropriate, for example the approach used in **prodlim** package. Without ties (between event times and censoring times) the two methods provide the same result. Our package allows the user to choose between one of the two methods.

Note that consistency of estimators (2), (3) and (4) is only guaranteed whenever $x + y$ is smaller than the upper bound of the support of the censoring time. The estimates produced via Bayes' theorem and Kaplan-Meier (`CKM`) may not produce a valid bivariate distribution since it does not guarantee that the bivariate distribution function is monotone. The problem can be explained by the fact that, as the conditioning set $T_1 \le x$ changes, the redistribution to the right of the probability mass associated with censored observations also changes. In contrast to the other two methods, the estimators based on Kaplan-Meier weights (`KMW` and `KMPW`) are monotonic (distribution) functions, in the sense that they attach positive mass to each observation.

2.5. Marginal distribution of the second gap time

From (1), (2), (3) and (4) we may obtain an estimator for the marginal distribution of the second gap time, $F_2(y) = P(T_2 \le y)$, namely

$$\widehat{F}_2(y) = \widehat{F}_{12}(+\infty, y) = \widehat{F}_1(+\infty)\widehat{F}_{KM}(y|\Delta_1 = 1) \tag{5}$$

$$\widetilde{F}_2(y) = \widetilde{F}_{12}(+\infty, y) = \sum_{i=1}^{n} W_i I(\widetilde{T}_{2i} \leq y) \tag{6}$$

Note that if $\widehat{F}_1(+\infty) = 1$, then (5) is the Kaplan-Meier estimator based on $(\widetilde{T}_{2i}, \Delta_{2i})$'s such that $\Delta_1 = 1$ (i.e., for which the first gap time is uncensored). Estimator (6) is different because the Kaplan-Meier weights W_i in this estimator are based on the $\widetilde{Y}_i$-ranks rather than on the $\widetilde{T}_{2i}$-ranks. In fact, since T_2 and C_2 are expected to be dependent, the ordinary Kaplan-Meier estimator of F_2 (estimator (5)) will be generally inconsistent. The corresponding estimator for (3) is obtained using the same ideas as for (6) by replacing the weights W_i by the presmoothed Kaplan-Meier weight $W_i^{\star}$ previously defined. Similarly, from Lin's estimator (4) one can obtain an estimator for the marginal distribution of the second gap time. Again, note that such estimator does not guarantee monoticity.

Results of an extensive simulation study comparing the four methods are reported in the paper by Meira-Machado and Moreira (2010). In this work the authors consider two simulation scenarios, the first scenario is the same as the described in Lin *et al.* (1999) (see their Section 3). In the second scenario, the gap times were generated using a family of bivariate Weibull distributions (see our Section 4 for more details). The main conclusions are the following: (a) the `CKM` estimator has larger bias for higher values of the first gap time; (b) the `KMW` estimator has less bias than its presmoothed version (`KMPW`); however, as expected, provides large standard errors in estimation (resulting in a wiggly estimator with fewer jumps); (c) the `KMW` and `IPCW` estimators are almost unbiased but the last one obtains higher levels of variance for small values of the second gap time.

Other estimators were proposed to estimate the bivariate distribution function. A valid estimator of the bivariate distribution function, was provided by Keilegom (2004) which is based on Akritas (1994). However, this approach has some limitations since some smoothing is required. Recently, alternative estimators for these quantities were also given in Keilegom, de Uña- Álvarez, and Meira-Machado (2011). This methodology assumes that the vector of gap times (T_1, T_2) satisfies the nonparametric location-scale regression model, allowing for the transfer of tail information from lightly censored areas to heavily ones.

3. Package description

The **survivalBIV** software contains functions that calculate estimates for the bivariate distribution function. As mentioned in Section 2, this package can be used to implement four methods (`CKM`, `KMW`, `KMPW` and `IPCW`). This software is intended to be used with the R statistical program R Development Core Team (2011). Our package is composed of 9 functions that allow users to obtain estimates for the bivariate distribution function. Table 1 provides a summary of the functions in this package.

Users can obtain the estimates for the methods discussed in Section 2 by means of three functions, namely, `survBIV`, `summary` and `plot`. Details on the usage of these functions can be obtained with the corresponding help pages. It should be noted that to implement the methods described in Section 2 one needs the following variables: `time1`, `event1`, `time2` and `event2`. Covariates have not been included in any of the implemented methods, therefore they are not necessary. The variable `time1` represents the observed time of the first event (first gap time), and `event1` the status indicator of the first gap time (if the first gap time

Function	Description
`dgpBIV`	A function that generates bivariate censored gap times from some known copula functions. By default returns a dataset of class `survBIV`.
`corrBIV`	Provides the correlation between the bivariate times for some copula distributions.
`survBIV`	Provides the adequate dataset for implementing all the four methods. The new dataset is of class `survBIV`.
`bivCKM`	Provides estimates for the bivariate distribution function for the conditional Kaplan-Meier estimator, `CKM`.
`bivIPCW`	Provides estimates for the bivariate distribution function for the inverse probability of censoring weighted estimator, `IPCW`.
`bivKMW`	Provides estimates for the bivariate distribution function for the Kaplan-Meier weighted estimator, `KMW`.
`bivKMPW`	Provides estimates for the bivariate distribution function for the Kaplan-Meier presmoothed weighted estimator, `KMPW`.
`plot`	A function that provides the plots for the bivariate distribution function and marginal distribution of the second time.
`summary`	Summary method for objects of class `survBIV`.

Table 1: Summary of functions in the package.

is a censored observation, the value is 0 and otherwise the value is 1). The variable `time2` represents the observed second time (second gap time). If `event1 = 0`, the second gap time is not observed and then `time2 = 0`. The variable `event2` is the final status of the individual (takes the value 1 if the second event of interest is observed and 0 otherwise).

4. Data generation

Users may use the function `dgpBIV` to generate bivariate survival data. This function can be used to generate bivariate survival times from two of the most known copula functions: Gumbel's bivariate exponential distribution Lu and Bhattacharya (1990, 1991), also known as the Farlie-Gumbel-Morgenstern distribution and the bivariate Weibull distribution. In the book by Johnson and Kotz (1972) several bivariate distributions are discussed and procedures of construction are given.

It is well known that Exponential and Weibull distributions are very useful for modeling survival times. The Farlie-Gumbel-Morgenstern distribution is given by $F(x,y) = F_1(x)F_2(y)[1+\delta(1-F_1(x))(1-F_2(y))]$ where the marginal distribution functions F_1 and F_2 are exponential with rate parameter θ_i, $i = 1,2$ and where $|\delta| \leq 1$ is the association parameter. The case of independence is obtained for $\delta = 0$ while the maximum of correlation (between T_1 and T_2) for the bivariate exponential distribution is obtained for $\delta = 1$ with bound equal to 0.25. These and other theoretical correlations between the bivariate times for this copula distribution (with unit marginal distributions) can be obtained using the input commands shown below.

```
R> library("survivalBIV")
R> corrBIV(dist = "exponential", corr = 0, dist.par = c(1, 1))
R> corrBIV(dist = "exponential", corr = 1, dist.par = c(1, 1))
```

In the following, using the `dgpBIV` function we will simulate bivariate exponential survival data (`dist = "exponential"`). We will use this data to explain and illustrate how numerical output for all methods can be obtained using the functions in the package. We will follow the simulation scenario described by Lin *et al.* (1999). We will simulate 1000 observations (`n = 1000`) assuming a maximum correlation of 0.25 (`corr = 1`) and use an independent uniform censoring time (`model.cens = "uniform"`), according to model $U(0, 3)$ (`cens.par = 3`).

```
R> set.seed(1500)
R> sim_data_exp <- dgpBIV(n = 1000, corr = 1, dist = "exponential",
+    model.cens = "uniform", cens.par = 3, dist.par = c(1, 1))
```

To obtain the estimates for the methods proposed in Section 2 we can use the functions shown in Table 1. As in the simulation by Lin *et al.* (1999) we are going to obtain estimates for bivariate distribution at values `t1 = 0.5108` and `t2 = 0.9163`. The true value is 0.2976. The following input command provides the estimate for the `KMW` method. With this command we obtain the pointwise confidence intervals (`conf = TRUE`) using a 1000 bootstrap replicates (`n.boot = 1000`). The construction of the pointwise confidence intervals is obtained by randomly sampling the n items from the original data set with replacement. This can be achieved using percentile bootstrap (`method.boot = "percentile"`) or using basic bootstrap (`method.boot = "basic"`). By default all functions use the `percentile` bootstrap (Davison and Hinkley 1997).

```
R> bivKMW(object = sim_data_exp, t1 = 0.5108, t2 = 0.9163, conf = TRUE,
+    conf.level = 0.95, n.boot = 1000)

               2.5%     97.5%
0.3015313 0.2692518 0.3337527
```

One important issue is whether 1000 is a suitable number of resamples to generate. Since a second and a third set of 1000 resamples gave similar results for the bootstrap confidence intervals, this suggests that with these number of resamples the results are consistent. From this perspective 1000 would seem sufficient.

The CPU time needed for running the `bivKMW` function varies according to whether bootstrap confidence bands are requested or not, the sample size, and the type of processor in the PC computer. The command presented above took no more than 2 second on a PC with an Intel Core i7 processor with 8 GB memory. The same input command but with n = 10000 resamples took a little more than 17 seconds.

Results for the other methods are very similar and can be obtained using the functions `bivKMPW`, `bivCKM` and `bivIPCW` with the same arguments. The `bivIPCW` function has one extra argument which allows the user to choose how to estimate $\widehat{G}$ in (4) `method.cens = "KM"` for the Kaplan-Meier method and `method.cens = "prodlim"` for the method proposed in **prodlim** package. In general, the two methods (for estimating the survival of censoring times) provide similar results; without ties (e.g., using simulated data) they provide the same result. The method based on Kaplan-Meier is implemented in C language and is faster.

The `summary` function can be used to obtain estimates for the bivariate distribution function. This function allows the user to obtain the estimates for all four methods using `method = "all"`:

```
R> summary(object = sim_data_exp, t1 = 0.5108, t2 = 0.9163, conf = TRUE,
+    conf.level = 0.95, n.boot = 1000, method = "all")

F( 0.5108 , 0.9163 )=
$CKM
               2.5%     97.5%
0.3001276 0.2695496 0.3321113
$IPCW
               2.5%     97.5%
0.2905816 0.2569556 0.3252051
$KMPW
               2.5%     97.5%
0.2982088 0.2669539 0.3274368
$KMW
               2.5%     97.5%
0.3015313 0.2688383 0.3338806
```

The CPU time needed for running the command presented below took a little more than 16 seconds. The same input command but with a sample size of $n = 10000$ took a little more than 68 seconds. Note that this input command is the one which requires more computational effort since all methods are implemented with bootstrap confidence bands (optional).

One limitation of the so-called Farlie-Gumbel-Morgenstern families of bivariate cdf's, is that the correlation of T_1 and T_2 can never exceed 1/3 (0.25 in the bivariate exponential distribution). The bivariate Weibull distribution allows for a larger correlation, which makes it superior to Gumbel's bivariate exponential. The `dgpBIV` function allows the user to generate a pair of times from the bivariate Weibull distribution with two-parameter marginal distributions. Its survival function is given by

$$S(x,y) = P(T_1 > x, T_2 > y) = exp\left[-\left[\left(\frac{x}{\theta_1}\right)^{\frac{\beta_1}{\delta}} + \left(\frac{y}{\theta_2}\right)^{\frac{\beta_2}{\delta}}\right]^{\delta}\right]$$

where $0 < \delta \leq 1$, and each marginal distribution has shape parameter β_i and a scale parameter θ_i, $i = 1, 2$. The correlation between the two gap times may be obtained though it is a complicated function of the shape and scale parameters and of δ. Again, the function `corrBIV`, from the **survivalBIV** package can be used to calculate the theoretical correlation between times for this bivariate distribution. This function may be valuable for choosing the appropriate shape and scale parameters. For example, choosing $\delta = 0.6$, $\theta_1 = \theta_2 = 7$ and shape parameters $\beta_1 = \beta_2 = 2$, lead to about 54% of correlation. Below follow two input commands to illustrate the use these two functions. The first command provides the theoretical correlation while the second generates bivariate survival data from the bivariate weibull with exponential censoring with rate parameter 0.08.

```
R> corrBIV(dist = "weibull", corr = 0.6, dist.par = c(2, 7, 2, 7))
R> sim_data_wei <- dgpBIV(n = 200, corr = 0.6, dist = "weibull",
+    model.cens = "exponential", cens.par = 0.08, dist.par = c(2 ,7 ,2 ,7),
+    to.data.frame = TRUE)
```

It is important to note that the conditional Kaplan-Meier estimator can be obtained using the **survival** (Therneau 2012) package alone. For example, for `t1 = 6.7006` and `t2 = 8.8805` this can be obtained through the following input commands:

```
R> library("survival")
R> KM1 <- survfit(Surv(time1, event1) ~ 1, data = sim_data_wei)
R> KM2 <- survfit(Surv(time2, event2) ~ 1, data = sim_data_wei,
+    subset = c(time1 <= 6.7006 & event1 == 1))
R> CKM <- (1 - summary(KM1, time = 6.7006)$surv) * (1 -
+    summary(KM2, time = 8.8805)$surv)
```

However, the `bivCKM` function in our package is simpler and allows the user to obtain the same estimate together with the bootstrap confidence bands:

```
R> sim_data_wei2 <- with(sim_data_wei, survBIV(time1, event1, time2, event2))
R> bivCKM(object = sim_data_wei2, t1 = 6.7006, t2 = 8.8805)
```

The **survival** package can also be used to obtain the marginal distribution of the second gap time for the `CKM` method. According to Equation 5, this can be obtained using the following input commands:

```
R> dft1 <- survfit(Surv(time1, event1) ~ 1, data = sim_data_wei)
R> dft2 <- survfit(Surv(time2, event2) ~ 1, data = sim_data_wei,
+    subset = (event1 == 1))
R> (1 - summary(dft2, time = 8.8805)$surv) * (1 -
+    summary(dft1, time = max(summary(dft1)$time))$surv)
```

Again, our package is simpler and it provides bootstrap confidence bands. Users can easily obtain these results for a specific method (using one of the four functions) or for all methods. The input commands are shown below.

```
R> bivCKM(object = sim_data_wei2, t1 = Inf, t2 = 8.8805, conf = TRUE,
+    conf.level = 0.95, n.boot = 1000)
R> summary(object = sim_data_wei2, t1 = Inf, t2 = 8.8805, conf = TRUE,
+    conf.level = 0.95, n.boot = 1000)
```

In addition to the numerical results graphical output can also be obtained. This will be shown in the next section using data from the well-known bladder cancer study. Details about this dataset are given below.

5. Example of application: Bladder cancer study

The methods described in Section 2 are illustrated using data from a bladder cancer study (Byar 1980) conducted by the Veterans Administration Cooperative Urological Research Group. In this study, patients had superficial bladder tumors that were removed by transurethral resection. Many patients had multiple recurrences (up to a maximum of 9) of tumors during the study, and new tumors were removed at each visit. For illustration purposes we re-analyze data from 85 individuals in the placebo and thiotepa treatment groups;

these data are available as part of the R **survival** package. Here, only the first two recurrence times (in months) and the corresponding gap times, T_1 and T_2, are considered. From the total of 85 patients, 47 relapsed at least once and, among these, 29 experienced a new recurrence. We have a total amount of censoring of 66% from which 44.7% is obtained from censored observations on the first gap time. We have about 38% of censored total time among the uncensored first gap time.

There is a high percentage of censored total time (Y's) which in general lead to difficulties in the estimation of the bivariate distribution function. The presence of a reasonable amount of censored Y's among the uncensored T_1's suggests that presmoothing could lead to an important reduction of variance in estimation (see de Uña-Álvarez and Amorim 2011).

We will calculate estimates for the bivariate distribution function in several points and plot these estimates. This will be done using the **survivalBIV** package.

In the following, we will demonstrate the package capabilities using data from the bladder cancer study. Below is an excerpt of the data with one row per individual.

```
R> data("bladderBIV", package = "survivalBIV")
R> head(bladderBIV)
```

```
 time1 event1 time2 event2
     1      0     0      0
     4      0     0      0
     7      0     0      0
    10      0     0      0
     6      1     4      0
    14      0     0      0
```

Each line represents the information from one individual in study. Among the first five observations, only individual represented by line 5 had a recurrence. This individual had a recurrence on month 6 and remained alive and without second recurrence until time 10 (months). Note that `event1 = 0` and `event2 = 0` (the remaining five observations) corresponds to a censored first gap time in the initial state ("remained alive without a recurrence"). All observations with `event1 = 1` and `event2 = 1` corresponds to individuals with a first recurrence and a second recurrence.

We computed the estimated values for the four estimators of $F_{12}(x, y)$, for x equals to 3, 13, 29 and 49 and y values 3, 10, 17.75 and 36.75, corresponding to marginal survival probabilities of 0.25, 0.5, 0.75 and 0.95. For illustration purposes we only report the estimated values of $F_{12}(x, y)$ for two pairs of gap times with 95% bootstrap confidence intervals.

```
R> bladder_obj <- with(bladderBIV, survBIV(time1, event1, time2, event2))
R> summary(object = bladder_obj, t1 = 13, t2 = 10, method = "all",
+    conf = TRUE, n.boot = 10000)
R> summary(object = bladder_obj, t1 = 29, t2 = 36.75, method = "all",
+    conf = TRUE, n.boot = 10000)
```

```
F( 13 , 10 )=
$CKM
```

```
                 2.5%      97.5%
0.16836961 0.08841834 0.25478264
$IPCW
                 2.5%      97.5%
0.15100626 0.06731521 0.24149771
$KMPW
                 2.5%      97.5%
0.16815396 0.09382032 0.25254201
$KMW
                 2.5%      97.5%
0.17192598 0.09163352 0.26381938

F( 29 , 36.75 )=
$CKM
               2.5%     97.5%
0.4498655 0.3276738 0.5755503
$IPCW
               2.5%     97.5%
0.4932222 0.3499283 0.6320595
$KMPW
               2.5%     97.5%
0.4303138 0.3090228 0.5603079
$KMW
               2.5%     97.5%
0.4349590 0.3119461 0.5603330
```

In this case it is clearly seen that the four methods can provide quite different results, specially for higher values of x or y (where the censoring effects are stronger). The CPU time needed for running the input commands presented above took no more than 2 minutes.

The outputs for the bivariate distribution function and for the marginal distribution of the second gap time are useful displays that greatly help to understand the patients course over time. Plots for these two quantities can easily be obtained. Figure 1 plots the marginal distribution function of the second gap time (time from first to second recurrence) for all methods. These plots are obtained using the following input commands:

```
R> plot(bladder_obj, plot.marginal = TRUE, method = "KMW",
+    ylim = c(0, 0.65), xlim = c(0, 45))
R> plot(bladder_obj, plot.marginal = TRUE, method = "KMPW",
+    ylim = c(0, 0.65), xlim = c(0, 45))
R> plot(bladder_obj, plot.marginal = TRUE, method = "IPCW",
+    ylim = c(0, 0.65), xlim = c(0, 45))
R> plot(bladder_obj, plot.marginal = TRUE, method = "CKM",
+    ylim = c(0, 0.65), xlim = c(0, 45))
```

In Figure 1 we can see new insights for each method, for example, about the number of jump points and monotonicity. In this graphical output we have on top the semiparametric estimator (right) and the method without presmoothing. The main difference between the

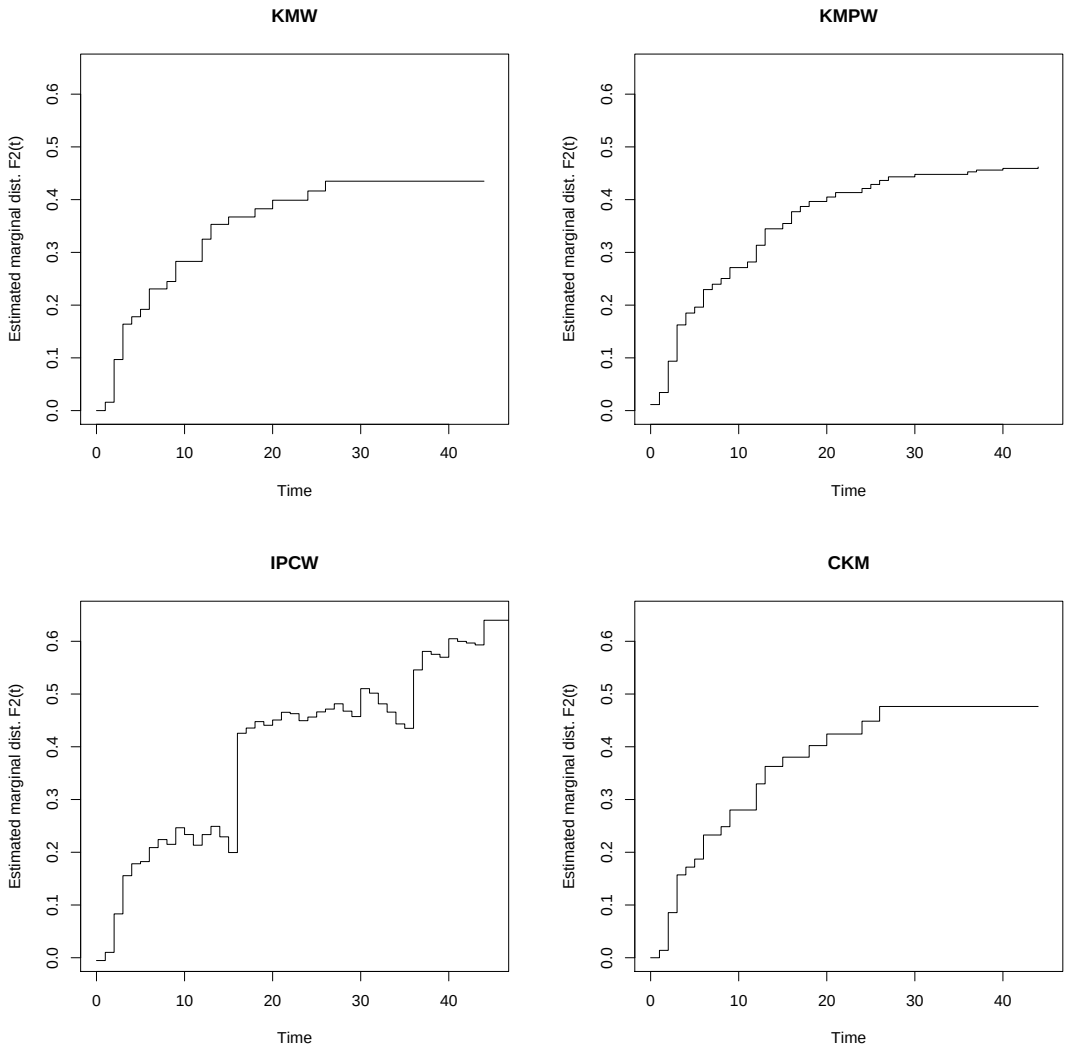

Figure 1: Marginal distribution function of the second gap time. Bladder cancer data.

first two methods is that the semiparametric estimator has more jump points, explicitly the censored values of the total time for which the first gap time is uncensored. Below, the method based on Bayes' theorem (`CKM`) and the method based on inverse censoring. Clearly, we can see that estimator based on inverse censoring (`IPCW`) provides a plot with more jump points than the remaining methods. Note also that this method provides non-monotone curves. In regard to the number of jump points and monotonicity, similar behaviors can be found in the plots for the bivariate distribution function (Figures 2 and 3). For illustration purposes we only present the plot for the semiparametric method. These plots are obtained through the following input command,

```
R> plot(bladder_obj, plot.bivariate = TRUE, method = "KMPW")
```

Plots for the different methods can be obtained by simply changing the `method` argument.

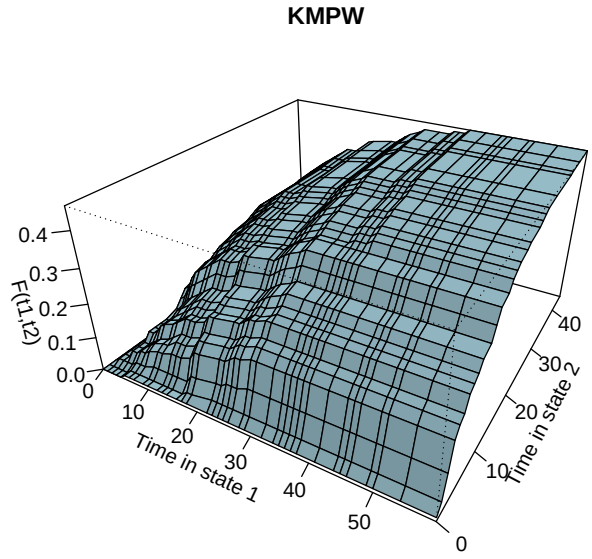

Figure 2: Bivariate distribution function. Bladder cancer data.

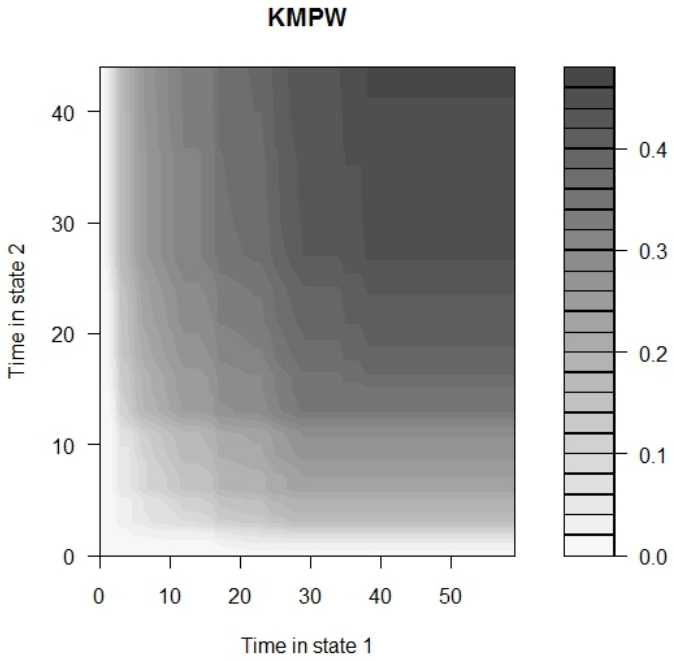

Figure 3: Contour plots for the bivariate distribution. Bladder cancer data.

6. Conclusion

This paper discusses implementation in R of some newly developed methods for the bivariate distribution function for censored gap times. The **survivalBIV** package uses four nonparametric and semiparametric estimators. One of these estimators is the conditional Kaplan-Meier, based on Bayes' theorem and Kaplan-Meier estimator; also, two recent estimators based on the Kaplan-Meier weights pertaining to the distribution of the total time (time to the second or final event of interest). It also implements the inverse probability of censoring weighted estimator proposed by Lin *et al.* (1999). The package allows for numerical results as well as graphics to be easily obtained. Covariates have not been included in our methods. This is a topic of current research and hopefully will be implemented in future. We plan to constantly update **survivalBIV** package to cope with other estimators.

Acknowledgments

The authors acknowledge receiving financial support from the Portuguese Ministry of Science, Technology and Higher Education in the form of grants PTDC/MAT/104879/2008 and SFRH/BD/62284/2009. This research was financed by FEDER Funds through "Programa Operacional Factores de Competitividade – COMPETE" and by Portuguese Funds through FCT – "Fundação para a Ciência e a Tecnologia", within the Project Est-C/MAT/UI0013/2011. The authors are grateful to the reviewers for their helpful comments.

References

Akritas MG (1994). "Nearest Neighbor Estimation of a Bivariate Distribution under Random Censoring." *The Annals of Statistics*, **22**, 1299–1327.

Andersen PK, Borgan Ø, Gill RD, Keiding N (1993). *Statistical Models Based on Counting Processes*. Springer-Verlag, New York.

Byar DP (1980). "Veterans Administration Study of Chemoprophylaxis for Recurrent Stage I Bladder Tumors: Comparisons of Placebo, Pyridoxine and Topical Thiotepa." *Bladder Tumors and Other Topics in Urological Oncology*, **18**, 363–370.

Davison AC, Hinkley DV (1997). *Bootstrap Methods and Their Application*. Cambridge University Press, New York.

de Uña-Álvarez J, Amorim AP (2011). "A Semiparametric Estimator of the Bivariate Distribution Function for Censored Gap Times." *Biometrical Journal*.

de Uña-Álvarez J, Meira-Machado LF (2008). "A Simple Estimator of the Bivariate Distribution Function for Censored Gap Times." *Statistics and Probability Letters*, **78**, 2440–2445.

Dikta G (1998). "On Semiparametric Random Censorship Models." *Journal of Statistical Planning and Inference*, **66**, 253–279.

Hougaard P (2000). *Analysis of Multivariate Survival Data*. Statistics for Biology and Health. Springer-Verlag, New York.

Johnson NL, Kotz S (1972). *Distributions in Statistics: Continuous Multivariate Distributions*. John Wiley & Sons, New York.

Kaplan EL, Meier P (1958). "Nonparametric Estimation from Incomplete Observations." *Journal of the American Statistical Association*, **53**, 457–481.

Keilegom IV (2004). "A Note on the Nonparametric Estimation of the Bivariate Distribution under Dependent Censoring." *Nonparametric Statistics*, **16**, 659–670.

Keilegom IV, de Uña- Álvarez J, Meira-Machado L (2011). "Nonparametric Location-Scale Models for Successive Survival Times under Dependent Censoring." *Journal of Statistical Planning and Inference*, **141**, 1118–1131.

Laan MJVD, Hubbard AE, Robins JM (2002). "Locally Efficient Estimation of a Multivariate Survival Function in Longitudinal Studies." *Journal of the American Statistical Association*, **97**, 494–507.

Lin DY, Sun W, Ying Z (1999). "Nonparametric Estimation of the Time Distributions for Serial Events with Censored Data." *Biometrika*, **86**, 59–70.

Lu JC, Bhattacharya GK (1990). "Some New Constructions of Bivariate Weibull Models." *Annals of Institute of Statistical Mathematics*, **42**, 543–559.

Lu JC, Bhattacharya GK (1991). "Inference Procedures for a Bivariate Exponential Model of Gumbel Based on Life Test of Component and System." *Journal of Statistical Planning and Inference*, **27**, 383–396.

Meira-Machado L, de Uña-Álvarez J, Cadarso-Suárez C, Andersen PK (2009). "Multi-State Models for the Analysis of Time to Event Data." *Statistical Methods in Medical Research*, **18**, 195–222.

Meira-Machado L, Moreira A (2010). "Estimation of the Bivariate Distribution Function for Censored Gap Times." *Proceedings of the 19th International Conference on Computational Statistics*, pp. 1367–1374.

R Development Core Team (2011). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

Therneau T (2012). ***survival**: A Package for Survival Analysis in S*. R package version 2.36-12, URL http://CRAN.R-project.org/package=survival.

Affiliation:

Ana Cristina Moreira, Luís Meira-Machado
Department of Mathematics and Applications

University of Minho
4810-058 Azurém, Guimarães. Portugal
Telephone: +351/253510400
Fax: +351/253510401
E-mail: id2809@alunos.uminho.pt, lmachado@math.uminho.pt

Journal of Statistical Software http://www.jstatsoft.org/
published by the American Statistical Association http://www.amstat.org/
Volume 46, Issue 13 *Submitted:* 2011-01-27
March 2012 *Accepted:* 2011-12-15